UniPR: Unified Object-level Real-to-Sim Perception and Reconstruction from a Single Stereo Pair

Supplementary Material

6. Overview

We provide **videos** that offer a brief introduction to UniPR, along with additional visualizations on real-world data and real-robot experiments. All video files are included in the Supplementary Material.

The formulation and implementation details of the Pose-Aware Shape VAE are provided in Sec. 7. Similarly, further details of the LVS6D dataset are described in Sec. 8. More ablation studies are provided in Sec. 9. With the emergence of SAM 3D Objects, we also clarify the conceptual and technical differences between their approach and UniPR, and present several comparative examples in Sec. 10. Real-world experimental results of UniPR are presented in Sec. 11. For a more comprehensive understanding of reconstruction quality, we include visual comparisons between Coders [40] and UniPR in Sec. 12. Moreover, qualitative results on the proposed LVS6D dataset are summarized in Sec. 13. Finally, detailed evaluation metrics and visualizations on public datasets are provided in Sec. 14.

7. Details for PASR Design

Details for Pose-Aware Shape VAE Our VAE model is based on 3DShape2VecSet [39] and takes object surface point clouds in an object-centered coordinate system as input. We adopt a spherical voxel space by sampling query points from a unit sphere rather than a cube, which is more suitable for representing objects under diverse rotations. The objective of our VAE model is to encode objects into lightweight embeddings for use in the detection pipeline. All ground-truth shape distributions for detection pipeline are generated using the encoder of our pretrained VAE model.

For the encoder $\mathcal{E}_{\text{VAE}}$ we first utilize self-attention layers to extract information from $\boldsymbol{z}_{\text{surface}} \in \mathbb{R}^{N\times C}$ to point latent embedding $\boldsymbol{z}_{\text{point}} \in \mathbb{R}^{N\times C}$.

$$\boldsymbol{z}_{\text{point}} = \text{SelfAttn}(\boldsymbol{z}_{\text{surface}}) \tag{10}$$

We initialize the object latent embedding, $\boldsymbol{z}_{\text{object}} \in \mathbb{R}^{C}$, and employ cross-attention layers to learn a mapping from the point distribution to the object embedding, as described in the main text. To ensure an effective representation and reduce the latent channel, we apply KL regularization during this process.

For the decoder $\mathcal{D}_{\text{VAE}}$, we first generate the sampled object latent embedding $\boldsymbol{z}_{\text{sampled}} \in \mathbb{R}^{C}$, as outlined in the main text. Next, we initialize the point latent embedding, $\hat{\boldsymbol{z}}_{\text{point}} \in \mathbb{R}^{N\times C}$, using a standard Gaussian distribution. Cross-attention layers are then utilized to recover information from the sampled object latent embedding $\boldsymbol{z}_{\text{sampled}}$ into the point latent embedding:

$$\hat{\boldsymbol{z}}_{\text{point}} = \text{CrossAttn}(\hat{\boldsymbol{z}}_{\text{point}}, \boldsymbol{z}_{\text{sampled}}) \tag{11}$$

Further architectural details of the proposed pose-aware shape VAE are provided in the main text.

Generative performance of PASR. We evaluate the generative performance of PASR across various shapes within the same category, diverse object orientations, and its interpolation capabilities in both pose and shape, as illustrated in Fig. 6. The results demonstrate the robustness of the learned PASR embedding space.

Texture Retrieval. Our method focuses primarily on shape reconstruction and does not include a dedicated texture generation module. However, given the VecSet representation produced by UniPR, we can seamlessly integrate external texture generation models. In particular, we employ Hunyuan3D-Paint-v2.1 to synthesize high-quality textures based on our reconstructed shapes, as demonstrated in the teaser, Fig. 8 and Fig. 9.

8. Details for LVS6D

Dataset structure. Our dataset is organized using a folder-based structure. For each stereo image, we provide object masks along with comprehensive annotations, including object category, position, scale, rotation, and the corresponding 3D shape. The object shape data include textured meshes sourced from OmniObject3D [33] and Google Scanned Objects (GSO) [5]. OmniObject3D contains 6,000 scanned objects across 190 daily-use categories, while GSO provides over 1,000 high-quality 3D-scanned household items.

The proposed LVS6D dataset spans 192 categories and includes more than 6,300 scanned objects. We generate approximately 0.4M stereo images for training and 1,000 images for testing, using over 500 high–dynamic-range (HDRI) backgrounds. Rendering is performed with BlenderProc 2.6.1 [4]. The stereo camera configuration uses a 13,cm baseline, and all images are captured at a resolution of 1920×1200 pixels.

In addition to synthetic data, we also capture several real-scene images using the identical stereo setup to evaluate real-world generalization. Overall, LVS6D is, to our knowledge, the largest stereo category-level object dataset designed for 6D pose estimation and shape reconstruction to date.

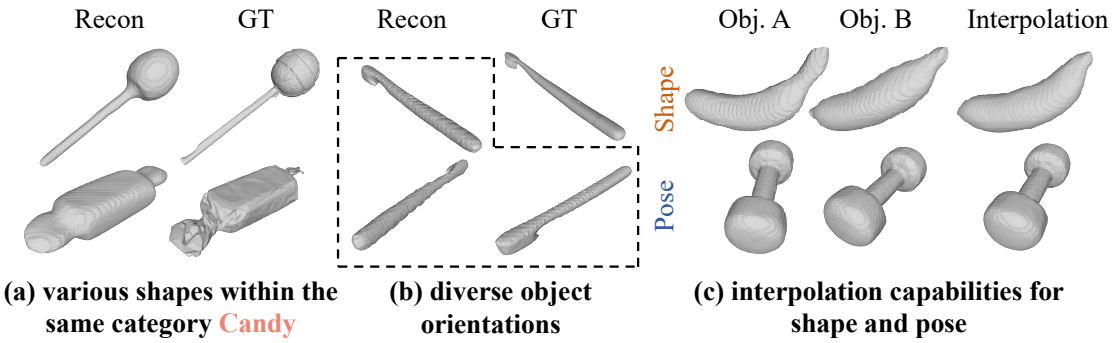

Figure 6. **Generative performance of PASR.** We evaluate PASR's generative performance on same-category shapes (a), diverse object orientations (b), and pose-shape interpolation capabilities (c). Results demonstrate the robustness of the learned embedding space.

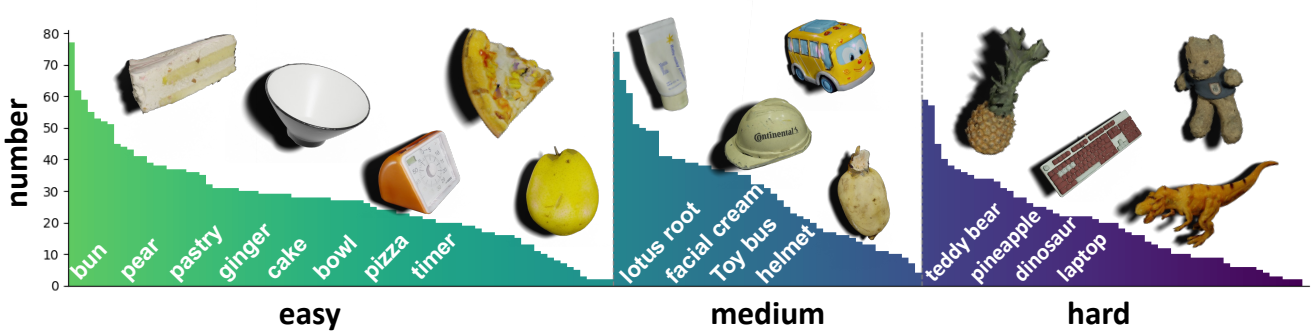

Figure 7. **Category Distribution of the LVS6D dataset.** We divide the dataset into three subsets Easy, Medium, and Hard based on the reconstruction difficulty. As the difficulty increases, the objects in each subset exhibit more complex geometries.

LVS6D subsets. As discussed in the main text, we divide the categories in LVS6D into Easy, Medium, and Hard subsets according to object complexity. Tab. 6 lists ten representative categories for each subset. Objects in the Easy subset exhibit simple and regular geometries (e.g., spheres, cubes) and display relatively low intra-class variation. In contrast, the Medium and Hard subsets contain objects with increasingly complex shapes and significantly higher intra-class variability, as shown in Fig. 7. For all evaluation metrics, we compute the results for every category individually and report the mean values for each subset.

9. Additional Experimental Results

9.1. Ablation Study on Loss Function

The loss term $\mathcal{L}_{\text{shape}}$ comprises both the KL divergence loss $\mathcal{L}_{\text{kl}}$ and the reconstruction loss $\mathcal{L}_{\text{recon}}$, as defined in Eq.7. We also conducted an ablation study without KL-based supervision, as shown in Tab. 7. The results demonstrate the importance of KL-based supervision for utilizing the pre-trained VAE model.

9.2. Fairness Comparison on Input Modalities

When comparing stereo methods with monocular baselines like Trellis, a potential concern is the inherent scale ambiguity present in monocular inputs. To verify that UniPR's performance gains stem primarily from architectural innovation rather than merely the sensor modality, we conducted extended experiments across four configurations: monocular Trellis, Trellis-stereo (extending Trellis with stereo input using ground-truth crops), Ours-mono (our pipeline adapted for monocular input), and our full stereo UniPR.

As shown in Tab. 8, UniPR-mono significantly outperforms both Trellis variants in geometric consistency (CD and F-Score). Furthermore, our full stereo model exceeds Trellis-stereo by a wide margin in Shape Proportion Error (SPE). This confirms that our proposed Pose-Aware Shape Representation and the end-to-end Triplane integration are the primary drivers of the observed improvements, demonstrating that UniPR efficiently fuses geometric constraints to achieve superior metric-scale reconstruction.

9.3. Part-Aware Refinement on TOD

In evaluating our method on the TOD dataset, we observed a performance discrepancy in categories with com-

Subsets	Object Samples
Easy	peach, ball, cake, bread, bun, egg, medicine bottle, mango, pomegranate, walnut
Medium	biscuit, candy, cheese, chili, doll, lotus root, small box, shoe, soap, teapot
Hard	pineapple, dinosaur, razor, teddy bear, keyboard, laptop, bamboo shoots, banana, clock, glasses case

Table 6. **Representative categories of LVS6D across subsets.** Ten representative categories are listed per subset: Easy subset objects have simple regular geometries, while Medium/Hard subsets include increasingly complex shapes and higher intra-class variability.

Method	AP↑	APE↓	ACD↓
w.o. KL-based supervision	0.675	1.210	2.947
Ours	**0.752**	**1.248**	**1.224**

Table 7. **The ablation of KL-based supervision.** The results demonstrate the importance of KL-based supervision for utilizing the pretrained VAE model.

Method	Input Modality	CD ↓	F-Score ↑	SPE ↓
Trellis	Monocular	0.1096	0.334	0.475
Trellis-stereo	Stereo (GT Crops)	0.1001	0.388	0.376
Ours-mono	Monocular	0.0141	0.868	0.110
Ours (Full)	**Stereo**	**0.0083**	**0.883**	**0.109**

Table 8. **Fairness comparison on input modalities and reconstruction metrics.** Our method demonstrates superior geometric consistency and shape proportion accuracy compared to Trellis, regardless of the input modality.

plex topologies, such as the "mug" category. The primary cause of this gap is that baseline methods like Coders [40] employ a part-aware prediction strategy to determine rotation based on specific features (e.g., handle positioning). In contrast, our original pipeline relied on global orientation without explicit part-level priors. Furthermore, another primary driver of the observed performance gap is that the baseline models are pre-trained on the comprehensive SS3D dataset, whereas our model is trained from scratch.

To ensure a rigorous and fair comparison, we implemented a part-aware refinement stage specifically for mug-like objects. As shown in Tab. 9, this enhancement enables UniPR to effectively resolve topological complexities and outperform Coders across most overall evaluation metrics.

Method	Refinement	Mug		Overall	
		5° 2cm ↑	10° 5cm ↑	5° 2cm ↑	10° 5cm ↑
Coders	Part-Aware	**56.2**	81.9	64.8	90.8
Ours (Original)	Global-only	36.8	74.6	63.2	86.9
Ours (**Enhanced**)	**Part-Aware**	46.6	**95.0**	**69.1**	**97.5**

Table 9. **Evaluation on the TOD dataset with part refinement.** By incorporating part-aware refinement, our enhanced model successfully resolves topological complexities and maintains its superiority in real-to-sim perception tasks.

10. Comparison with SAM 3D Objects

SAM 3D Objects is a recent image-to-3D object generation pipeline developed concurrently with UniPR. Here, we clarify the key differences.

First, although SAM 3D unifies segmentation and reconstruction, it still relies on an external detection module for object perception. In contrast, UniPR provides a fully end-to-end perception–reconstruction pipeline without requiring separate detection or segmentation stages.

Second, UniPR leverages stereo vision, whereas SAM 3D operates on monocular images. Stereo geometry provides metric depth cues, enabling accurate real-to-sim transfer and supporting real-world robotic grasping experiments, as demonstrated in Sec. 11. This depth-aware design also leads to more reliable shape proportions in many cases and allows UniPR to handle objects with ambiguous or hard-to-recognize categories (e.g., stones and unstructured shapes) as shown in Fig. 8, Fig. 9 and Fig. 10.

Third, SAM 3D processes objects individually, requiring one-by-one reconstruction. UniPR, by contrast, performs parallel full-scene processing in a single forward pass, offering significantly higher efficiency and better utilization of global context.

In summary, UniPR differs fundamentally from SAM 3D Objects. Rather than serving as a general-purpose image-to-3D generator, UniPR targets tabletop object perception, metric reconstruction, and downstream robotic manipulation, making it a specialized and practically oriented model within the real-to-sim domain.

11. Results on Real-world Data

We further present qualitative comparisons on real-world scenes in Fig. 12, where we compare UniPR with Coders. The results demonstrate that UniPR exhibits strong generalization capability when applied to real-world inputs.

To further validate the metric accuracy of UniPR's predicted object positions and scales, we conduct real-robot experiments using a simple top-to-bottom grasping policy. By directly using UniPR's metric-scale pose and shape predictions as input to the grasping policy, the robot successfully grasps a variety of objects, confirming the practical reliability of our method. Grasping results are provided in the accompanying video.

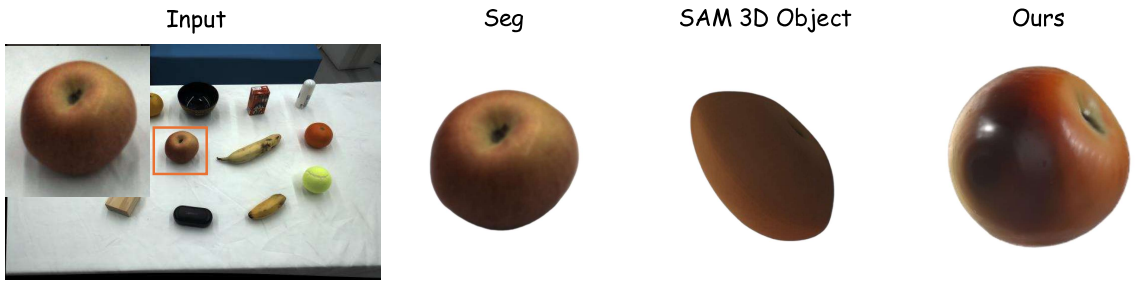

Figure 8. Comparision with SAM 3D Object on real scene.

Figure 9. Comparision with SAM 3D Object on simulated scene.

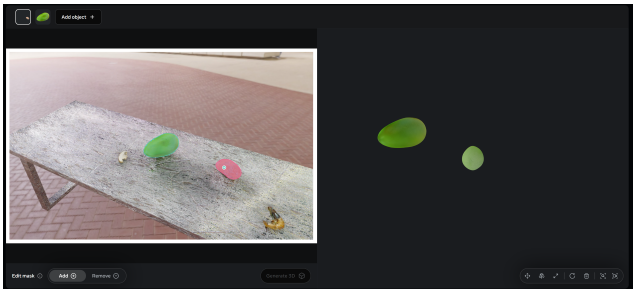

Figure 10. Illustration of SAM 3D Object test result.

12. Visualization of Reconstruction Results

We present the visualization of reconstruction results on the LVS6D dataset in Fig. 13. As demonstrated, our UniPR method consistently produces high-quality meshes, showcasing the effectiveness of pose-aware shape reconstruction. Our UniPR does not need category-level prior so we can generate a more detailed object mesh.

13. Qualitative Results on LVS6D

We present qualitative results on LVS6D in Fig. 11. Notably, the rotation predictions produced by UniPR demonstrate significant improvements compared to UniPR. In addition, UniPR achieves superior performance in mesh reconstruction.

14. Detailed evaluation on public datasets

We present a detailed evaluation of the 16 categories in the SS3D dataset which validates the effectiveness of our method. Additionally, we visualize the detection results in Fig. 14. Furthermore, Tab. 10 provides a detailed analysis of our method's performance on TOD.

Category	Banana	Bool	Bottle	Bowl	Carrot	Cork	Cucumber	Cup
Coders-3D_{50}	72	90	71	59	42	89	34	62
Ours-3D_{50}	**96**$_{+24}$	**97**$_{+7}$	**80**$_{+9}$	**98**$_{+39}$	**88**$_{+46}$	**97**$_{+8}$	**95**$_{+61}$	**98**$_{+36}$
Coders-3D_{75}	26	51	16	20	11	45	9	16
Ours-3D_{75}	**60**$_{+34}$	**83**$_{+32}$	**54**$_{+38}$	**63**$_{+37}$	**59**$_{+48}$	**69**$_{+24}$	**57**$_{+48}$	**59**$_{+43}$
Coders-5°5cm	47	65	**80**	58	51	74	38	72
Ours-5°5cm	**75**$_{+28}$	**87**$_{+22}$	75$_{-5}$	**96**$_{+38}$	**88**$_{+37}$	**90**$_{+16}$	**84**$_{+46}$	**96**$_{+24}$
Coders-5°2cm	26	43	33	23	22	48	20	33
Ours-5°2cm	**34**$_{+8}$	**48**$_{+5}$	**39**$_{+6}$	**63**$_{+40}$	**53**$_{+31}$	**49**$_{+1}$	**40**$_{+20}$	**61**$_{+28}$

Category	Dish	Fork	Knife	LargeBox	Orange	SmallBox	Scissors	Spoon
Coders-3D_{50}	88	54	**71**	52	35	66	48	42
Ours-3D_{50}	**95**$_{+7}$	**80**$_{+26}$	68$_{-3}$	**90**$_{+38}$	**95**$_{+60}$	**94**$_{+28}$	**94**$_{+46}$	**55**$_{+13}$
Coders-3D_{75}	37	15	12	9	6	32	18	10
Ours-3D_{75}	**56**$_{+19}$	**36**$_{+21}$	**45**$_{+33}$	**69**$_{+60}$	**41**$_{+35}$	**55**$_{+23}$	**50**$_{+32}$	**29**$_{+19}$
Coders-5°5cm	64	58	61	16	56	32	51	33
Ours-5°5cm	**79**$_{+15}$	**70**$_{+12}$	**62**$_{+1}$	**69**$_{+53}$	**95**$_{+39}$	**57**$_{+25}$	**77**$_{+26}$	**44**$_{+11}$
Coders-5°2cm	20	33	**33**	5	25	24	29	**20**
Ours-5°2cm	**37**$_{+17}$	**41**$_{+8}$	31$_{-2}$	**26**$_{+21}$	**52**$_{+27}$	**26**$_{+2}$	**33**$_{+4}$	19$_{-1}$

Table 10. **Detailed Results on the SS3D Test Dataset.** We evaluate UniPR on the SS3D test dataset with 16 categories of unseen objects. The first row corresponds to categories in reference to SS3D. Our method can manage objects across all 16 categories, encompassing various sizes, shapes and materials, demonstrating the generalization capability of our stereo framework. The gray rows in the table indicate results from our method and the numbers in the table represent accuracy percentages.

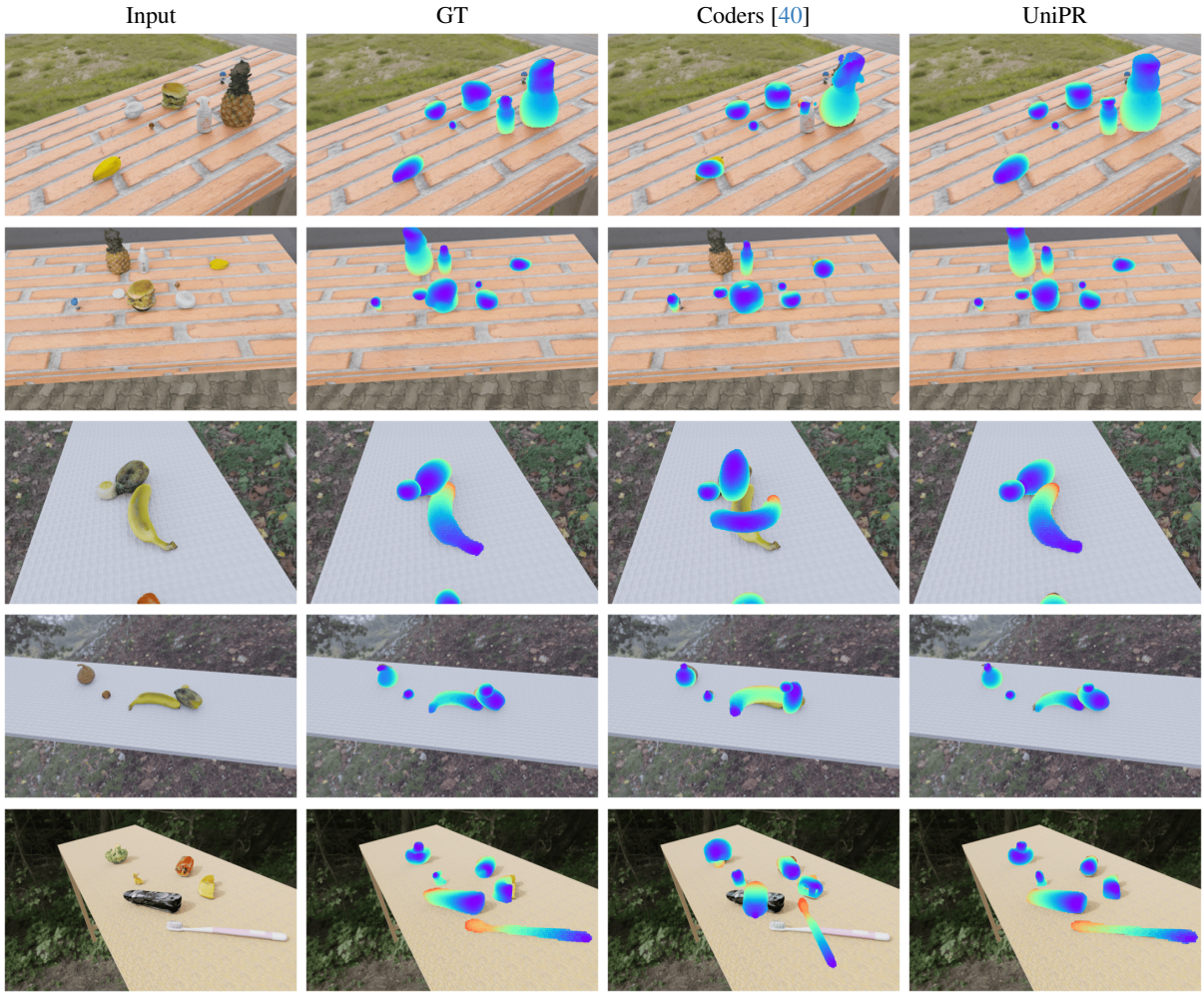

Figure 11. **Qualitative results on LVS6D dataset.** We present visualizations of Coders [40] and UniPR on the LVS6D dataset. The results highlight our superior performance on the LVS6D dataset.

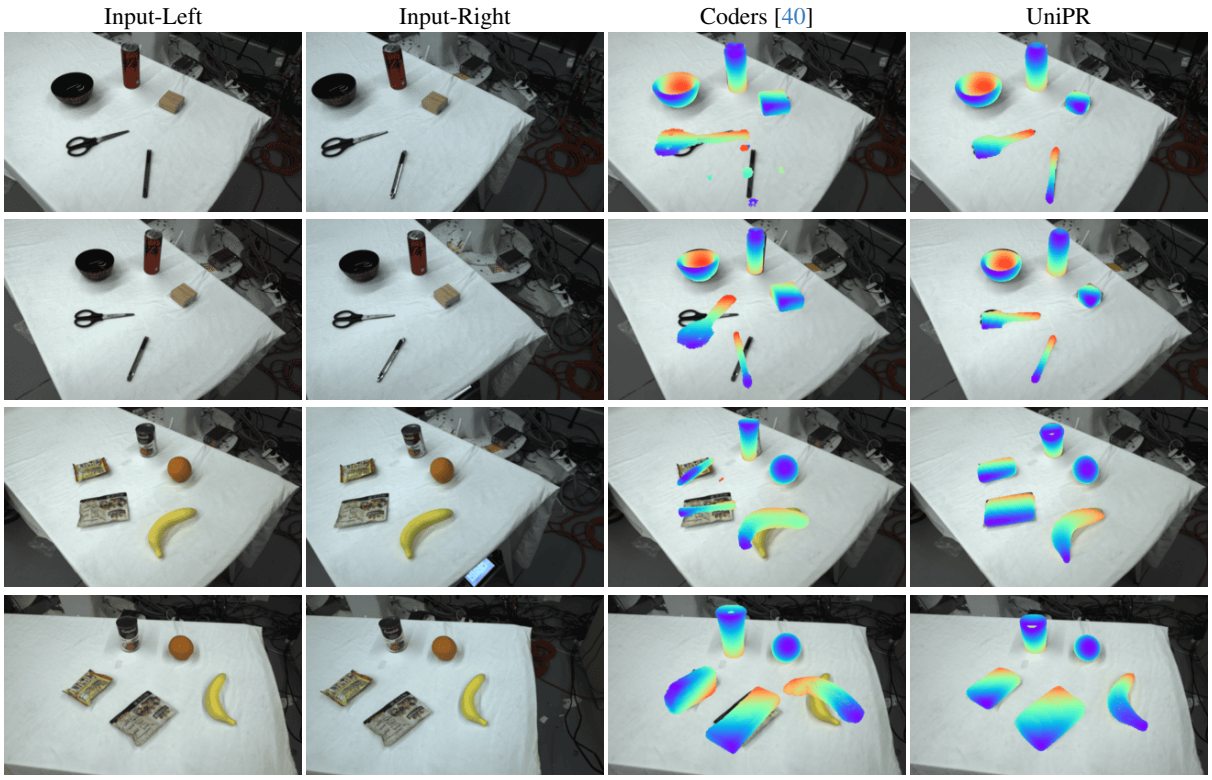

Figure 12. **Qualitative results on real-world data.** We present visualizations of Coders and UniPR on real-world data. We compare the UniPR with Coders [40]. The results demonstrate UniPR's strong generalization ability with real-world data.

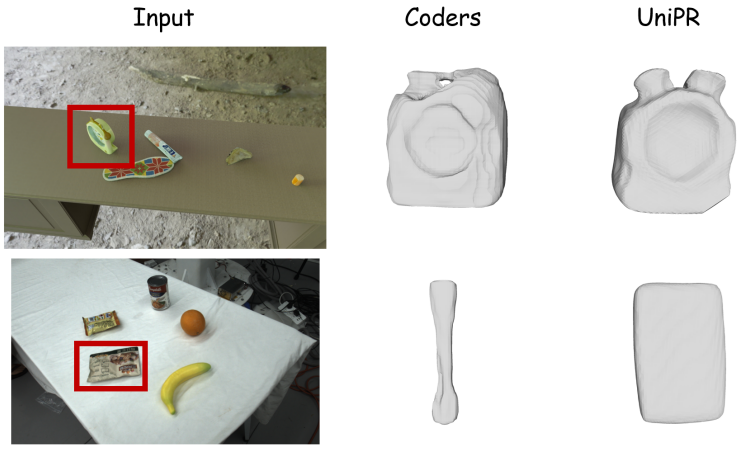

Figure 13. **Qualitative Results of Reconstruction.** Our approach generates meshes with high quality which is attributed to the effectiveness of pose-aware shape representation.

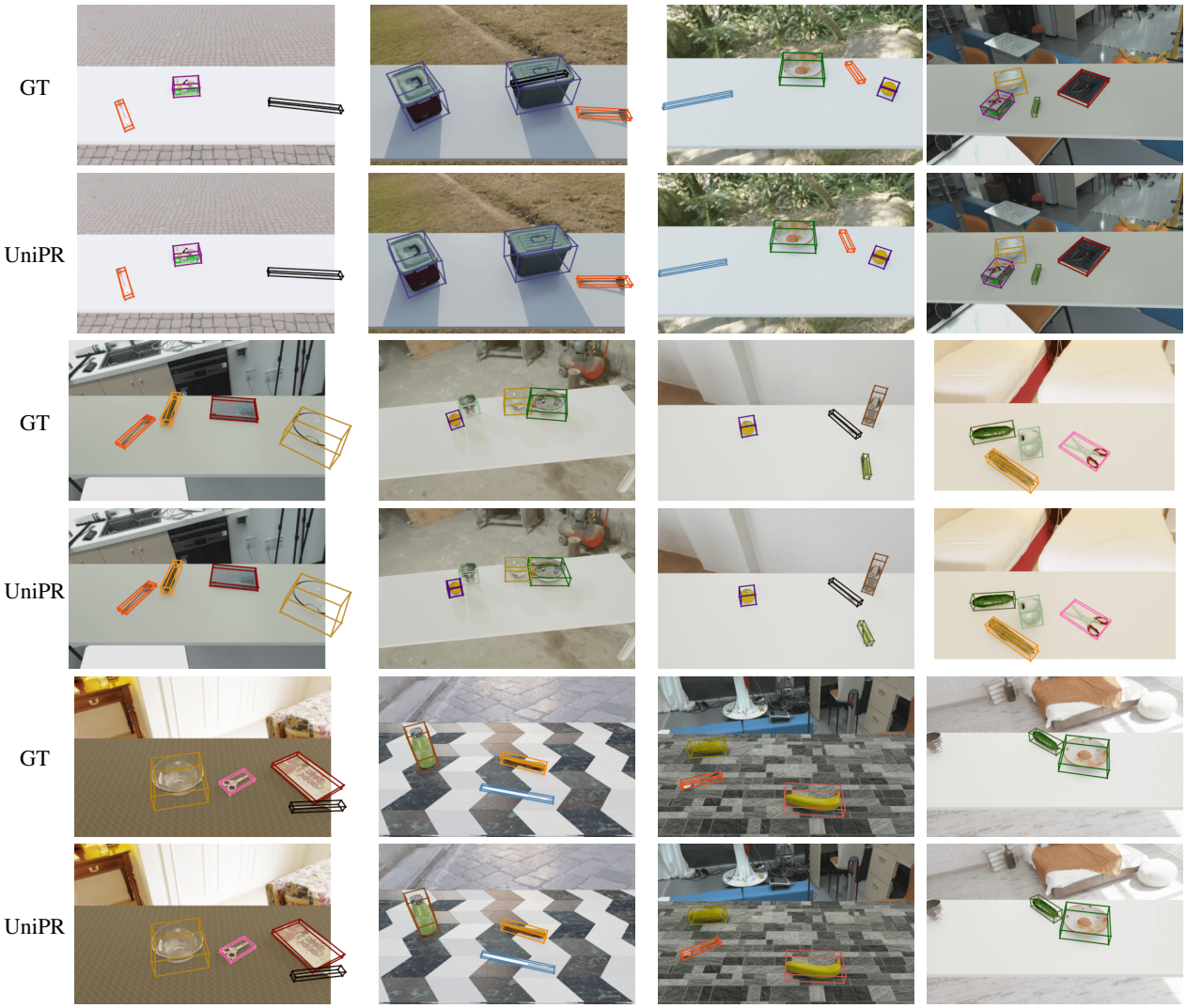

Figure 14. **Visualization of StereoPose on SS3D dataset.** Our proposed UniPR demonstrates consistently strong performance across various scenarios.